Appropriately Representing Military Tasks for Human-Machine Teaming Research

Chad C. Tossell(✉), Boyoung Kim, Bianca Donadio, Ewart J. de Visser, Ryan Holec, and Elizabeth Phillips

Warfighter Effectiveness Research Center (DFBL), USAF Academy,
Colorado Springs, CO 80840, USA
Chad.tossell@usafa.edu

Abstract. The use of simulation has become a popular way to develop knowledge and skills in aviation, medicine, and several other domains. Given the promise of human-robot teaming in many of these same contexts, the amount of research in human-autonomy teaming has increased over the last decade. The United States Air Force Academy (USAFA), for example, has developed several testbeds to explore human-autonomy teaming in and out of the laboratory. Fidelity requirements have been carefully established in order to assess important factors in line with the goals of the research. This paper describes how appropriate fidelity is established across a range of human-autonomy research objectives. We provide descriptions of testbeds ranging from robots in the laboratory to higher-fidelity flight simulations and real-world driving. We conclude with a description and guideline for selecting appropriate levels of fidelity given a research objective in human-machine teaming research.

Keywords: Human-machine teaming · Autonomy · Robots · Simulation · Fidelity

1 Introduction

Machines already play an integral role in defense. Across military services, operators team with machines to perform important tasks such as diffusing explosive ordinance or safely flying an aircraft. As automated technologies have advanced, so has the military's reliance on them. For example, remotely piloted aircraft (RPAs) extend the reach of humans and have kept numerous pilots out of harm's way. Newer automated systems are being fielded at higher levels of automation [1]. Instead of relying on pilots to shift to autopilot, the Auto Ground Collision Avoidance System (Auto-GCAS) on many fighter aircraft take control of the aircraft based on its own calculations. If the system detects the distance to ground and trajectory of the aircraft are unsafe, it will take control of the aircraft from the pilot and fly to safer airspace. Auto-GCAS has already been credited with saving seven lives and is a good example of how automation can team effectively with humans [2].

C. Stephanidis et al. (Eds.): HCII 2020, LNCS 12428, pp. 245–265, 2020.
https://doi.org/10.1007/978-3-030-59990-4_19

Autonomous systems have also been utilized in defense since World War II [3]. Autonomous systems differ from automated systems because autonomous systems can independently determine courses of action based on their knowledge of itself and the environment [4, 5] whereas automated systems are more restricted to execution of a set of scripted pre-determined sets of actions. Autonomous systems are expected to use this knowledge to achieve goals in situations that are not pre-programmed. Simply because they can operate independently in these unanticipated environments does not mean they will operate independent of humans. Indeed, most concepts of operations for military autonomous systems have these systems teamed with humans across warfighting and peacetime contexts. AI-based systems are already used in drone swarms to learn patterns based on their observation. Other forecasted applications of AI that are currently in development include search and rescue techniques and exoskeleton suits [6]. AI technologies will continue to penetrate battlefields to help human operators perceive complex battlespaces, fight effectively, and stay safer across military domains. One challenge to this end is to enable human-autonomy interactions that are effective and natural across a wide range of military tasks. Studies to improve trust, shared situational awareness, social norms, and collaboration in human-autonomy systems are actively being conducted to facilitate these interactions [7]. The United States Air Force (USAF), in particular, has provided a vision for humans teaming with autonomy [8]:

> *In this vision of the future, autonomous systems will be designed to serve as a part of a collaborative team with airmen. Flexible autonomy will allow the control of tasks, functions, sub-systems, and even entire vehicles to pass back and forth over time between the airman and the autonomous system, as needed to succeed under changing circumstances. Many functions will be supported at varying levels of autonomy, from fully manual, to recommendations for decision aiding, to human-on-the-loop supervisory control of an autonomous system, to one that operates fully autonomously with no human intervention at all. The airman will be able to make informed choices about where and when to invoke autonomy based on considerations of trust, the ability to verify its operations, the level of risk and risk mitigation available for a particular operation, the operational need for the autonomy, and the degree to which the system supports the needed partnership with the airman. In certain limited cases the system may allow the autonomy to take over automatically from the airman, when timelines are very short for example, or when loss of lives are imminent. However, human decision making for the exercise of force with weapon systems is a fundamental requirement, in keeping with Department of Defense directives.*

This paper describes the approach being used to help develop these capabilities within the Warfighter Effectiveness Research Center (WERC) at the United States Air Force Academy (USAFA). Realistic simulations have been developed to explore trust, workload, human-robot interaction (HRI), social norms, and performance across a wide range of military tasks. Importantly, the fidelity requirements of the intelligent agents, tasks, and environments have been carefully designed according to research goals and applications to future battlefields.

2 Fidelity in Human-Machine Teaming Research

Cadets studying at USAFA are involved in this research as part of the research team and as participants for experiments. As part of the research team, select firstie (i.e., senior) cadets majoring in human factors (HF) engineering and behavioral sciences learn important concepts in HF (e.g., trust, workload, and situation awareness [9, 10]) by helping to design experiments in HMT. Another learning goal is to afford opportunities for these cadets to critically think about how intelligent systems might be involved in future warfighting. Cadets learn quickly that the effective integration of these systems is important for military tasks with life-or-death consequences. Thus, in determining fidelity requirements for HMT studies, we oftentimes aim for higher degrees of realism.

To achieve this goal, different levels of fidelity have been used in research environments to represent a variety of Air Force tasks. Fidelity has been defined as the degree to which a simulation replicates reality [11]. When a simulation more closely mimics the real world, it is higher fidelity. Simulations lower in fidelity are more artificial without as many matching elements in the real world. In medical and aviation training, high-fidelity simulations have included full-body manikins programmed to provide realistic physiological responses to care and 360°, full-motion flight simulators to prepare pilots for live flight. Low-fidelity trainers in the same domains include patient vignettes read from a sheet of paper to test medical students and chair flying [12, 13]. Fidelity in these contexts have generally referred to elements of the simulation environment (e.g., graphics, haptics, etc.) and labeled *physical fidelity* [14].

Fidelity has also been characterized beyond simply the physical features of the environment. To create immersive experiences in gaming environments and realistic behaviors in psychological experiments, fidelity has been considered based on human elements. *Conceptual fidelity* measures the degree to which the narrative/scenario elements in a simulation are connected and make sense to humans in the loop. Similarly, *psychological/emotional fidelity* is the extent to which the task mimics the real-world task to provide a sense of realism [12]. *Cognitive fidelity* has also been used to measure the level of human engagement with simulations [15]. Fidelity has thus been more broadly defined to capture the extent to which the environments and other elements of the simulation come together to elicit the intended emotional, cognitive, and behavioral responses from humans. Even simulations low in physical fidelity can create visceral human reactions and realistic responses to stimuli (e.g., crying in response to reading sad vignettes on a piece of paper [16]). Games can provide humans a very immersive experience when the narrative, gameplay, and graphics are combined in consistent ways and not based on synchronization or the level of physical fidelity [17–24].

Experiments in HMT, like most psychology experiments, are (among other things) designed to examine specific questions of interest by measuring effects through partitioning the sources of variability. Studies using artefacts that are low in physical fidelity have been able to examine antecedents and consequences of mind perception toward robots. Tanibe and colleagues read vignettes of robots sustaining damage in a kitchen to participants before obtaining perceptions of mind via questionnaires [27]. High-fidelity simulations have also been used in HMT research. In an elegantly designed study of mistrust, Alan Wagner had participants work in a building where a simulated fire with smoke started. Participants were then instructed to follow a robot out of the building.

Even though the robot navigated incorrectly to a room with no exit (by design), participants still frequently followed the robot. Participants seemed to experience real risk and relied on the robot despite it making an obvious error. Studies across the fidelity spectrum have attempted to maximize internal, external, and/or ecological validity to test research hypotheses and improve the precision of the results.

In HMT studies at USAFA, we have set up controlled, yet realistic, environments to maximize internal, external, and ecological validity. Cadets, many of whom will be users of highly automated and autonomous systems, are the participants in these studies. The simulations used in our experimental settings were developed to maximize the human-centered types of fidelity at low cost. We have used well established methods such as Wizard of Oz (WoZ) for both high-fidelity and low-fidelity simulations to try and mimic future functions of machine agents [28, 29]. Additionally, we have developed novel tasks in higher-fidelity settings to study trust in autonomous tools (e.g., a Tesla). Across these experimental settings, cadet researchers work with military operators and faculty members to study HMT in future warfighting scenarios [25]. We describe a subset of these settings along with examples of studies below (Table 1).

Table 1. Research testbeds described in this paper

Testbed	Simulation technology	Ecological target	Example studies
A. Autonomous Flight Teaming (AFT)	F-35 Flight Sim (Prepar3d, COTS)	Flying an F-35 with 4 autonomous F-16s	Workload on Trust, SA in Multitasking in Air Combat [30]
B. Human Automation Research in a Tesla (HART)	Tesla Model X	Trust in auto systems in risky tasks	Trust in Risky Autopark [31–33]
C. Gaming Research Laboratory	Games (e.g., Overcooked)	AI agent teaming in interdependent tasks	Teaming with AI vs Human on Task Performance [26]
D. Social Robotics Laboratory	Robot APIs (Pepper, NAO, Aibo)	Robotic teammates using natural language	Social norms with robots [34, 35]

3 Human-Machine Teaming Research Settings at USAFA

3.1 Testbed A: Autonomous Flight Teaming (AFT)

As mentioned above, one vision for future flight operations is for autonomous aircraft to seamlessly integrate with human F-35 pilots [8]. To explore factors and evaluate designs to facilitate this integration, we adapted a flight simulator to allow human pilots to fly with virtual autonomous F-16s (Fig. 1). The AFT consists of three features that must work together to provide an environment for human participants to operate their simulated aircraft in a team of autonomous systems: Hardware/software, flight scenarios, and measurement systems.

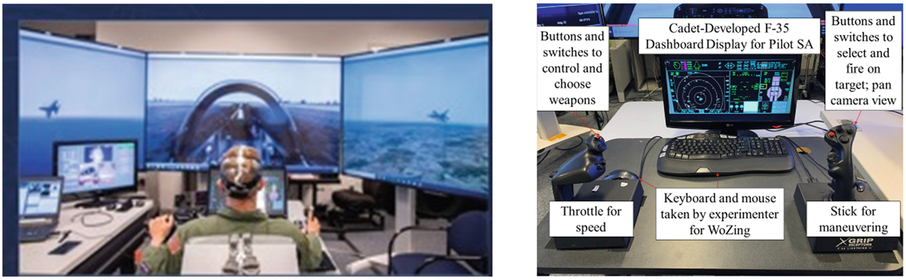
Buttons and switches to control and choose weapons
Cadet-Developed F-35 Dashboard Display for Pilot SA
Buttons and switches to select and fire on target; pan camera view
Throttle for speed
Keyboard and mouse taken by experimenter for WoZing
Stick for maneuvering

Fig. 1. AFT displays and controls

3.1.1 Hardware and Software for Visualization and Control

As shown in Fig. 1, the AFT was designed as a high-fidelity simulator to mimic future autonomous flight teaming. Participants fly in the simulator using the F-35 Hand on Throttle and Stick (HOTAS), which is similar to the stick and throttle used in the actual F-35 Lightning II. The displays found in F-35 simulators used in USAF training are overlaid with new interfaces and models designed by cadets with input from subject matter experts (SMEs) and faculty at USAFA.

The flight simulator integrates three, 72-in. monitors into the testing environment with an additional visual display for prototyping an F-35 pilot's dashboard display. The integration of the four monitors into the testing environment, each with their own information stream, allows for more robust cognitive fidelity in the experiment that reflects the heavy workload and strained SA of actual pilots. A VR system can also be integrated into the AFT to further immerse the participant in the scenario or to study Wizard of Oz (WoZ) teaming scenarios.

Other artifacts are also integrated on a case-by-case basis. For example, pilots in live flight are continually checking their kneeboard and monitoring the gauges, radar screens, and the Heads-Up Display (HUD). The SMEs helped develop a "cheat sheet" that is used to mimic a pilot's kneeboard. This kneeboard includes the names of the targets at each SAM site, how to make a radio call, the call-signs of the autonomous F-16s, and the scenario-dependent flight parameters.

3.1.2 Scenario Development

Like gaming systems, our scenario development focused on blending the narrative, graphical elements, and physics of the simulation to create an immersive experience for participants. Our goal is in these scenarios is to replicate future operations in autonomous flight to a level where participants are highly engaged and motivated to succeed with their virtual autonomous F-16s. The general sequence is outlined here:

1. After pre-brief, take off from airfield and identify the SAM sites
2. Conduct an orientation flight to learn basic flight skills
3. Determine how to destroy SAM sites: self and/or autonomous F-16s
4. Neutralize targets and be on lookout for other threats
5. Have F-16s form up on F-35 to complete mission

The pre-mission brief establishes the importance of the mission and how participants can do well. Most of our studies do not offer real incentives for performing well in our scenarios. However, we attempt to increase motivation through artificial incentives (e.g., “to succeed in this mission, all surface-to-air missiles must be destroyed without any losses to your flight team”). We rely on cadets’ competitiveness in these activities to study ways they trust their autonomous wingmen.

Following the pre-brief, participants fly in three different scenarios; first, a familiarization scenario, followed by two operational scenarios. The familiarization scenario introduces the participant to the information streams of the four screens and guides them through using each of the flight controls. The participants also practice making radio calls and learn how to engage their autonomous F-16s based on the goals of the study. For example, one study assessed three different ways to communicate with the autonomous wingmen using supervisory control methods and a “play calling” technique [10].

With input from SMEs (i.e., experienced Air Force pilots), we have developed a range of scenarios at different levels of difficulty and workload in order to assess different ways participants trust, communicate, and team with autonomous wingmen. One scenario requires participants to attack enemy sites that each contain many surface-to-air missiles (SAMs). The participant leads all three of the autonomous F-16s in this mission. The participant can engage each of the targets individually (and likely fail) or rely on the autonomous wingmen to assist. New displays (e.g., Fig. 2) and methods to communicate (e.g., voice versus supervisory control on a gaming controller). Workload and difficulty levels are increased systematically by introducing air threats and/or increasing radio traffic. When air threats are introduced, participants must multitask with their autonomous wingmen to neutralize the air threat in addition to targeting ground threats with a limited number of missiles.

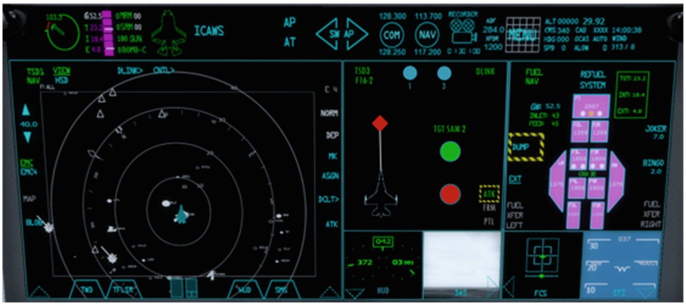

Fig. 2. Visual display that shows statuses of F-16s.

3.1.3 Measurement

We have examined questions involving workload, trust, and performance in teaming with autonomous systems via passive means using physiological sensors, telemetric performance measures, and experimenter/SME observations (Table 2). To avoid disrupting the scenario and influencing the immersive experience, the scenarios are not interrupted for measurements. The Tobii Pro Glasses system is worn like any other pair of glasses, and collects data on eye fixation, saccade patterns, and pupil dilation. This eye-tracking system provides the research team with clear behavioral metrics of the participant's attention, which research suggests can be linked to trust [9].

Additionally, overall team performance is assessed based on mission success (e.g., number of enemy targets destroyed, number of hits received, etc.). The built-in telemetry system allows for real-time recording of flight data (e.g., missiles fired, successful hits, location of wingmen, etc.), which augments behavioral observations and allows for quantifiable performance metrics (including their adherence to the flight parameters and time on target) to be analyzed post-experiment. Depending on the study, other measures can be collected (Table 2).

Table 2. Examples of data collected in the AFT.

Type	Construct	Metric
Behavioral	Performance, Attention	Telemetry, Eye-Tracking
Physiological	Cognitive Workload, Stress	EEG, ECG & GSR
Subjective	Workload, SA, Trust/Self-Confidence	TLX, SART, Post-Q
Observed	Performance	Mission Success Rate (# of Targets Destroyed)

3.1.4 Testbed Validation

Validation of the AFT has been conducted in proof-of-concept studies [30]. In a recent unpublished study, eight cadets engaged in risky behaviors because of the lack of real-world consequences. They relied less on their teammates and more on themselves and explicitly acknowledged flying too low or too high, aggressively going after SAMs, and performing advanced maneuvers during the mission that were not necessary. Thus, even though our simulation was higher in physical fidelity than other laboratory-based multitasking tasks (e.g., MAT-B), participants still recognized it was an artificial environment and their behaviors were reported as unrealistic.

The risky behaviors were minimized in follow-on (also unpublished) studies using higher-fidelity scenarios. Adding tasks such as keeping an altitude, not exceeding airspeed levels, and neutralizing air threats reduced the frequency of observed risky behaviors. Our current study (on hold due to COVID-19) is evaluating trust and performance as a function of workload and experience level. Across scenarios and experience levels, we are seeing variance in reliance on the autonomous wingmen and overall performance. Our measures are sensitive to the changing dynamics of the scenario and trust levels. For example, at the individual level, EEG has been captured and correlated to task load and trust levels (Fig. 3). All measures are time-synchronized and correlated for a more complete understanding of stress, workload, and SA. Even though there are no real consequences for mission failure, our goal of creating an environment to study future teaming concepts, trust, SA, and other phenomena in USAF tasks has provided an important look into HMT beyond more basic laboratory tasks.

3.2 Testbed B: Human Automation Research in a Tesla (HART)

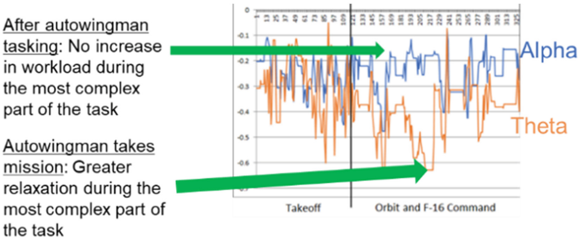

Fig. 3. Alpha and theta waves recorded from a participant during takeoff and initial task allocations with autonomous wingmen.

Fig. 4. The Tesla Model X - Air Force Version

Given the lack of real consequences in testbeds such as AFT, the HART testbed has allowed us to evaluate trust in real-world and potentially risky environments. While driving is obviously a different task than flying, there are similarities in trusting autonomous systems in both environments. The goal for HMTs is to engender calibrated trust where the expected performance of automation matches the actual performance of automation [36, 37]. Research has shown that humans may have a propensity to over-trust robots in realistic emergency scenarios [38]. This over-trust could lead an individual to underestimate the risk associated with using an intelligent agent or even foster misplaced reliance on technological teammates [39]. While there has been recent work attempting to develop an "adaptive trust calibration" system, which would help with issues of over-trust and under-trust [40], such a system has not been used and tested on a variety of technological agents in high risk environments with a focus on high physical and cognitive fidelity.

To address the need for assessing trust calibration in high risk environments, the WERC has established a mobile research laboratory known as HART (Human-Automation Research in a Tesla) mobile lab. This mobile lab environment is set up in a 2017 Tesla Model X (Fig. 4), equipped with various automated features which include lane-following, adaptive cruise control (ACC), and automated parking. Within the HART mobile lab are five distinct pieces of technology for data recording that do not impede the participant's experience, therefore maintaining psychological fidelity during the task. Unifying all this technology is its mobility, which allows researchers to collect a multitude of data in the most ecologically valid way in a dynamic, naturalistic environment. Like the AFT, we can collect data throughout the study on eye fixations via eye-tracking, stress fluctuations via GSR, and heart rate readings via ECG with the use of the Tobii Pro Glasses system and NeuroTechnology's BioRadio. The third piece of technology used in the HART is the Advanced Brain Monitoring B-Alert X24 Mobile electroencephalography (EEG). This EEG will allow us to measure workload and attention [41]. Additionally, cameras mounted inside the car will capture the interior and exterior environment but also the participant's face to analyze their emotional and cognitive states based on Ekman and Friesen [42] action unit measurements of facial muscle movement. The final piece of technology being used is a RaceCapture telemetry system, which will allow for the real-time recording of vehicle data (i.e. acceleration, braking, and steering).

3.2.1 Testbed Validation

We have conducted a series of studies to examine how trust in real autonomous systems develops [31–33]. An initial study evaluated driver intervention behaviors during an autonomous parking task (Fig. 5). While recent research has explored the use of autonomous features in self-driving cars, none has focused on autonomous parking. Recent incidents and research have demonstrated that drivers sometimes use autonomous features in unexpected and harmful ways.

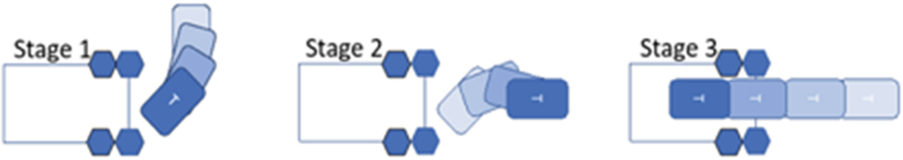

Fig. 5. The three distinct stages of the Tesla's autoparking feature. The Tesla is represented by the solid blue rectangles with the "T" on them. Time is represented by those rectangles transitioning from a lighter to a darker shade of blue. The trash cans are represented by the hexagons. (Color figure online)

Participants completed a series of autonomous parking trials with a Tesla Model X and their behavioral interventions were recorded. Participants also completed a risk-taking behavior test and a post-experiment questionnaire which contained, amongst other measures, questions about trust in the system, likelihood of using the autopark feature, and preference for either the autonomous parking feature or self-parking. Initial intervention rates were over 50%, but declined steeply in later trials (Fig. 6). Responses to open-ended questions revealed that once participants understood what the system was doing, they were much more likely to trust it. Trust in the autonomous parking feature was predicted by a model including risk-taking behaviors, self-confidence, self-reported number of errors committed by the Tesla and the proportion of trials in which the driver intervened. Using autonomy with little knowledge of its workings can lead to high degree of initial distrust and disuse. Repeated exposure of autonomous features to drivers can greatly increase their use. Simple tutorials and brief explanations of the workings of autonomous features may greatly improve trust in the system when drivers are first introduced to autonomous systems.

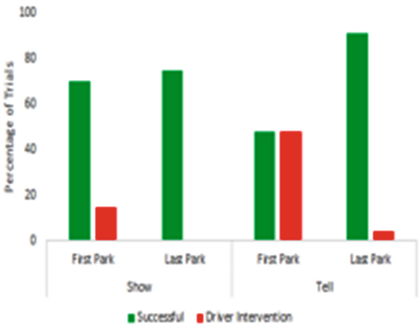

Fig. 6. Percentage of trials for which the autoparking was engaged, where the car was able to successfully and fully park itself, and where the driver intervened, split by first and last park, as well as condition.

In a follow-on study, we compared driver intervention rates when either showing the autoparking capabilities to drivers or merely telling them about the features without demonstration. The study showed that the intervention rates when showing the parking features drops significantly compared to when drivers are merely told about the autopark capabilities.

3.3 Testbed C: Gaming Research Laboratory at USAFA

3.3.1 AI Teaming Research

One issue that has emerged in HMT research is that human team players communicate less overall with autonomous teammates, which can affect team performance [43, 44]. This may be part of the reason that few AI agents exist that work with humans in a real-world team setting with the ability to communicate with human team members. There is thus a need to evaluate human communication styles with autonomous agents in team environments.

3.3.2 Gaming as a Reasonable and Fun Way to Approximate Teamwork

To study human-AI communication, we have established a video game laboratory to leverage the immersive experiences games afford to many people. Within this game laboratory, we developed a testbed called Cooking with Humans and Autonomy in Overcooked! 2 for studying Performance and Teaming (CHAOPT) [26]. Overcooked 2 is particularly immersive. The game requires coordinated teamwork and good communication to be successful (e.g., earn more points, advance to higher levels, etc.). The game uniquely, cleverly and dynamically manipulates the environment which forces flexible allocation of roles and information sharing (Fig. 7). We have added Wizard of Oz capabilities to observe how human behavior changes when a person believes they are working with an AI agent. This WoZ capability has been valuable because we can examine how humans react differently based on their teammate, with an emphasis on how they communicate, how they evaluate their teammate, and how much they trust their teammates.

3.3.3 Overcooked 2 Game Mechanics

Overcooked 2 is a video game that is teamwork intensive and requires communication to succeed at a high level. Players are given food orders to complete and are required to navigate the kitchen environment, prepare the orders, and deal with distracting features. Once orders are submitted, players are awarded points and tips based on the order correctness and priority. If orders are not fulfilled in time, that order goes away and players lose points and bonuses. Points, bonuses, completion time, and other game-based scores provide handy performance measures for studies. Players have tasks to

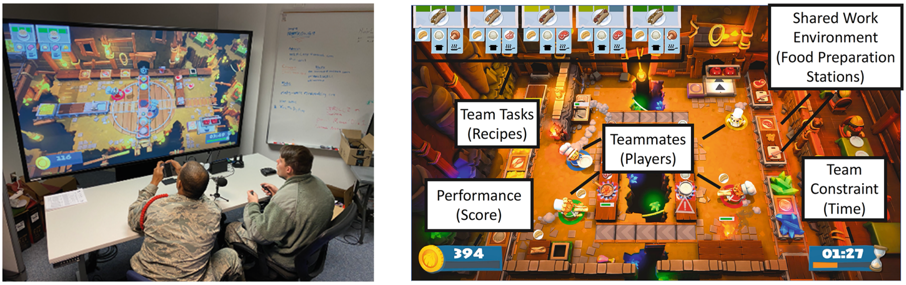

Fig. 7. Two cadets playing the Overcooked 2 video game in the Gaming Research Lab along with the game display to show how in-game tasks map to higher-level teaming concepts.

perform such as: collect the required order ingredients, chop food, clean dishes, and cook food to complete orders.

Overcooked stimulates communication through various tasks and levels of difficulty. Different worlds and levels introduce environments where a single player cannot complete the level on their own, which is where the communication aspect of this game is extremely vital. The beginning levels do not require as much communication as the player is still being introduced to new game concepts, but around World 2 Level 1, communication becomes more important to the team's success. The kitchen maps become restrictive and the supply locations become limited to all players. Communication becomes required to fulfill orders and complete the level. The methods of communication that are being measured in this pilot test are push versus pull communication and the amount of communication used. All communication is verbal for this experiment. Push communication refers to one player telling the other player what they want or need. For example, one player that needs cheese to complete an order may tell the other player, "Chop me cheese" or "Throw me cheese". Pull communication refers to one player asking for another player to complete a task. For example, the player may say "Can you throw me cheese?" or "Can you chop cheese for this order?" The communication method may indicate the level of trust the participant has in the human or autonomous agent or which confederate they would prefer to play with.

3.4 Testbed D: Social Robotics Laboratory

In future battlefields, some AI systems will likely be embodied in physical robots and not simply exist in virtual environments. Thus, human-robot interaction (HRI) environments must be developed to explore their coordination with humans. Like above, we have developed a testbed to mimic future environments to explore robots in authority, teaming alongside humans, as a moral advisor, and a teammate in a stressful set of tasks.

3.4.1 Warrior-Robot Relations

Using a commercially-available Pepper robot, we created a task to explore empathy towards robots in a human-robot team task under stress. In the transition of robots and other artificial agents from *tools* to *teammates* [45], important questions are raised by forming teammate like relationships with non-human agents, especially in battlefield environments. For instance, in military contexts, there have been powerful stories of military members disusing robots as a result of feeling too much empathy towards them. In 2007, an Army Colonel deeply empathized with an improvised explosive device (IED) detecting robot. The robot was designed to use its many articulating legs to purposefully detonate IEDs, and as a result sacrifice itself in the detonation process. The Army Colonel stopped an exercise demonstrating the robot sacrificing itself to detonate the IEDs, stating that it would be inhumane to continue the exercise [46]. Such inappropriate attributions of empathy towards robots and other artificial agents could prevent humans from making full use of the benefits of autonomous systems deployed in dangerous environments. If a human feels too much empathy, they may sacrifice elements of the mission or task to prevent harm to an autonomous teammate, which could jeopardize the mission or the team.

However, the ability to form empathetic bonds towards others is a major point of emphasis for the training of U.S. military officers especially regarding facilitating respect for the human dignity of others. Empathy may be needed to truly think of and rely upon artificial agents that are specifically designed to be teammates for humans. If a human does not feel enough empathy, the operator may not fully utilize the agent as a teammate or use it in a manner that is unintended or unsustainable. Thus, a balance between too little and too much empathy towards a robot will likely be needed to facilitate good teaming between humans and robots in military contexts. Initial evidence supporting the need for such a balance was provided by a set of studies, using the TEAMMATE simulation, that examined empathy for a robot in the context of a space mining mission [47, 48]. The studies demonstrated that human teammates responded more often and more quickly to a robot request for help when it was portrayed as a helpful teammate companion and when it appeared more damaged.

Additionally, military teams are normally required to operate under high levels of stress and stress is a common way comradery is built between members of the team. Inducing stress can help make people feel stronger perceived understanding of others' emotions and feelings [49]. As a result, interacting with a robot teammate under high stress could cause people to feel more empathetic towards a robot team member than when interacting under low stress.

To induce empathy and stress in a human-robot team task we conducted a study with cadets at USAFA. The task was to interact with Pepper, a humanoid robot developed by Softbank Robotics, as a teammate in an intrinsically incentivized spelling bee game. Participants were tasked with working with Pepper to spell increasingly difficult English words taken from the National Adult Spelling Bee Practice vocabulary list [50]. Participants were responsible for spelling 2/3 of the words, while Pepper was responsible for spelling 1/3 of the words. If participants spelled a word incorrectly, the Pepper would lose 1/8 of its simulated battery health/life. And, Pepper's simulated battery health and system performance would continue to degrade as the participant misspelled words until Pepper ran out of health. Pepper's speech was also slowed as a result of its diminished health. However, the participant could stop the study at any time to prevent Pepper from losing health. Upon doing so, the participant's spelling score would become final. If Pepper's simulated health was completely diminished, participants were told that Pepper lost all memory of the participant and was no longer functional.

As the participant interacted with Pepper, their poor spelling performance harmed Pepper by reducing Pepper's simulated battery life and health by 1/8. Participants' willingness to continue was perceived as less empathy shown towards Pepper. If/when the participant stopped the spelling task to preserve Pepper, it was recorded as an objective measure of empathy shown toward the robot. To induce stress, half of the participants completed the spelling task under time pressure.

Our results suggested an interesting interaction between participant gender and stress on empathy scores, where males and females showed a differential pattern of empathetic behavior toward the Pepper robot under different stress conditions [35]. The data trended such that males were initially more empathetic towards robot partners in unstressed conditions than females but were far more likely to act less empathetic toward robot partners in stressed conditions. Whereas females were far more likely to act more empathetic toward robot partners in stressed conditions. This finding could inform changes in the design of robots intended for stressful scenarios and when working in teams with people. For example, an engineer could design less human-like responses for women operators and more human-like responses for male operators based on the stressful dynamics of the task. However, it is important when designing human-like responses to adhere to intuitive principles of human-likeness [51] and avoid designs that are perceived as uncanny [52]. This work presents some early findings in understanding the nuanced roles that empathy and stress play in teaming with robotic and other artificial agents. This understanding will be important in creating effective human-agent teams for successful deployment in several Air Force contexts.

3.4.2 Robots as Moral Advisors

The role of an artificial moral advisor would be to assist people in making decisions that comply with moral standards and values [53–58] and perhaps even serve as a 'cognitive wingman' [59]. Previous research on how people make moral judgments and decisions showed that people in identical moral dilemmas may arrive at diverging decisions depending on various psychological factors, such as gender, time pressure, cognitive load, and language [60–63]. This inconsistency in people's decisions may become a critical problem, especially in contexts where their decisions could bring about significant and irrevocable consequences. One such example would be a military context such as the Air Force, where there is a high demand for

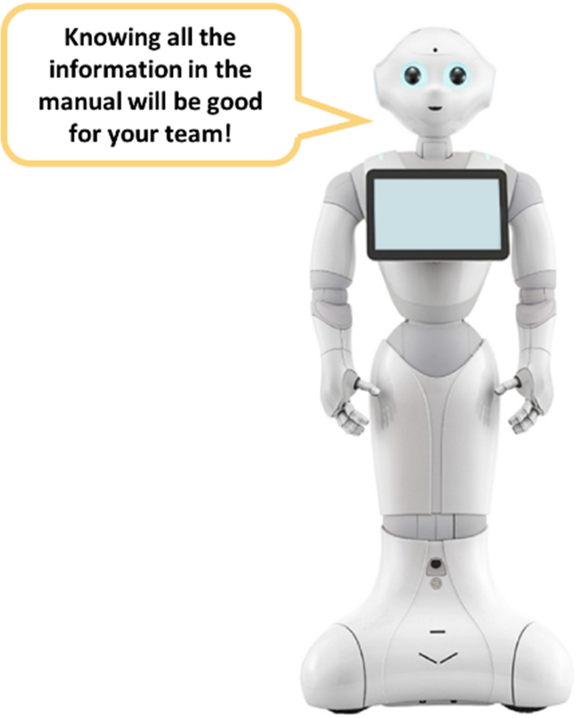
Knowing all the information in the manual will be good for your team!

Fig. 8. Pepper robot used in the social robotics task

morally-laden decisions [64]. To illustrate, imagine that the following ethical dilemma takes place in a military operation: An officer faces a decision of whether to sacrifice one person to save many lives or to take no action and lose many lives. Previous researchers found that, in this situation, the time it took for people to reach the same utilitarian decision of sacrificing one life to save many was approximately 5.8 s under no cognitive load but increased to 6.5 s under cognitive load [61]. This implies that in the military context, the officer's decision may cause strikingly different consequences depending on their cognitive resources available at the time. Given this volatile tendency of people's moral decision-making process, it would be useful to develop an artificial intelligence system that may guide human teammates to follow a systematic and well-informed decision-making process before reaching a moral decision. However, it is critical that such a system has a degree of moral competence [59, 65].

As a first step towards building an artificial moral advisor, we have launched an investigation on how robots can effectively communicate a message that encourages people to make morally right choices. Recently, it was found that a robot's responses to people's request can result in changes in people's judgments of whether a certain behavior is morally permissible or not [66]. If people's moral judgments could be shaped by a robot's response, would people also be receptive to a robot's unsolicited advice on what is morally right or wrong choice in human-robot teams? Drawing from Confucian role-ethics [67], we predict that a piece of moral advice from a robot may influence people's behavior to a varying degree depending on how people relate themselves with the robot. Whereas, in a human-robot team, a human may readily acquire a status of a partner, teammate, or colleague, a robot may not easily be granted with such a status [68]. Therefore, we predict that people would be more willing to follow a robot's moral advice when they perceive the robot as their teammate compared to when they do not perceive the robot as their teammate.

To test this idea, we will program commercial-off-the-shelf (COTS) robots, like pepper (Fig. 8), to interact with human participants. These participants are asked to do a tedious task that will ultimately benefit their team performance, but can stop at any time. In this situation, a moral behavior would be to complete the task. Whenever participants express their intention to stop continuing the task, a robot gives them a response that emphasizes the importance of being a good teammate. We predict that the effect of highlighting the importance of being a good teammate would have a more positive effect on the likelihood of participants' completing the task when they perceived the robot as their teammate compared to when they do not.

Our proposal to seek assistance from an artificial intelligence system in making moral decisions may raise many ethical concerns. Decisions about ethical dilemmas where no absolute answer is available may appear to be outside the purview of any kind of artificial intelligence systems. Perhaps, the value of an artificial moral advisor in human-machine teams may be most evident in situations where morally right or wrong actions are stipulated (e.g., "Do not cheat on tests"). However, even in moral gray areas, we expect that an artificial moral advisor can be useful at various levels of fidelity. First, an artificial moral advisor can quickly and accurately gather and convey to humans information relevant to decisions at hand so that humans can make informed decisions. Next, humans can verify whether their personal moral norms and values are consistent with their team's or the general public's by checking in with an artificial moral advisor. Third, humans can rely on an artificial moral advisor to facilitate clear discussions about moral choices with other human teammates. Finally, an artificial moral advisor can optimize persuasive strategies for encouraging humans to adhere to moral norms by accurately assessing team characteristics.

4 Discussion

There are many ways to define fidelity requirements for simulations used in HMT studies. Across the testbeds described above, a wide range of established and novel methods have been used to create presence and appropriate fidelity in research studies. These environments have elicited behaviors that appear natural and appropriate for the task. However, the results of these studies are intended to generalize to high stress environments where real lives are at stake. The primary question is whether we have achieved the appropriate level of fidelity that allows us to make this generalization. In applied studies when technologies are being tested, designing appropriate levels of fidelity to elicit stress and interaction become even more important. For example: How can researchers stimulate the same levels of stress as when a real air threat is detected in an F-35? We discuss these and other considerations in the next sections.

4.1 The Need for Appropriate Fidelity in Military Training and Research

The highest level of fidelity a researcher can have is to conduct research in actual operational and naturalistic environments. The use of systems in naturalistic environments can provide helpful data to inform lab-based studies. For example, effectiveness of early human-robot teamwork was demonstrated by operational robot deployment in the 9/11 World Trade Center attack and Fukushima Daiichi disasters [69]. Others have demonstrated the utility of interviewing pilots to discover how they learn their systems in situ. This information can be used in the design and develop automated systems such as the Auto-GCAS [2, 70].

However, conducting research in naturalistic settings poses its own set of challenges. Operational personnel are not always available for research, the occurrence of disasters cannot be predicted in advance, and naturalistic settings and conditions cannot usually not be fully controlled. This reality necessities the use of simulations. By developing high fidelity simulations in military contexts, the gap between developers and users can be bridged prior to production or fielding of new technology. For human-autonomy teaming research there are several unique considerations of fidelity requirements in the creation of such testbeds.

4.1.1 Simulating Stress, Risk and Urgency to Study Trust in HMT

The first consideration is the accurate simulation of stress, risk, and urgency. Military members rely on basic and advanced technologies in warfare. Establishing artificial time limits, promoting competition, and using the WoZ techniques may not be enough to create an appropriate sense of urgency. Additionally, IRB regulations require researchers to not put participants at more risk than they are exposed to daily. Researchers have addressed this issue by creating more realistic but still controllable environments. For example, trust in a robot was studied in a building simulated to be on fire with real alarms and simulated smoke [38, 39]. Our studies have used an actual Tesla vehicle with a realistic parking situation [31–33]. All these studies have seemed to elicit genuine behavior from participants making it more likely that constructs such as trust are evaluated more appropriately. However, it is unclear how these approximations differ from real life-or-death domains such as in medicine, aviation, and military operations. It is also unclear how different humans behave in these environments compared to laboratory studies. Further research into HMT fidelity requirements should compare different levels of fidelity to answer this question.

4.1.2 Using Wizard of Oz to Simulate Future HMT Capabilities

One difficult challenge in HMT research is the ability to simulate future HMT capabilities such as a fully autonomous agent that can communicate, coordinate and work seamlessly in an HMT. Currently, such an agent does not exist. Therefore, the WoZ method is a useful method to approximate this future capability. Participants usually participate in HRI/HCI studies tabula rasa and believe they are interacting with technology and not a human. Thus, we have found WoZ [28] to be a high-fidelity method to uncover differences in how participants interact with technologies versus humans. Such methods are used in the HRI field to uncover important psychological and human performance issues that might arise assuming the technology is realized. An alternative to this method is the “Oz-of-Wizard” in which the human behavior is simulated or assumed to test how a robot will respond [29]. Eventually, WoZ may become less important in research if autonomous agents become more readily available and configurable. For example, recent work in our laboratory has investigated bonding and trust with Sony’s Aibo, a robotic dog that demonstrates a variety of autonomous dog-like behaviors [71, 72].

4.2 The Trade-off Space for Fidelity Requirements

Fidelity is multi-faceted and certain tradeoffs should be made while maintaining the integrity of the training exercise, task assessment, or research study, depending on the core reason for using a simulation environment.

The goals an organization has for using a simulator (whether it be as a testbed for training, assessment, or a means of studying trust and teamwork) determine the fidelity requirements for that simulation environment. Therefore, a crucial step in implementing simulation practices is clearly identifying the needs or research focus of that initiative. For example, to get the most realistic results for measuring risk taking and trust in automated tasks, we were able to use the HART. Meanwhile, to understand how pilots may behave with systems that have autonomous capabilities which do not yet exist, in scenarios that are impossible to test in a naturalistic environment, we use the AFT, prioritizing physical and conceptual fidelity over psychological fidelity. Whereas, when testing teamwork and moral decision making, we value cognitive or conceptual fidelity above physical fidelity.

In summary, a brief guideline to assist in selecting the appropriate level of fidelity might be as follows:

1) Clearly identify the research objective for use of simulation
2) Based on the identified research objective, appropriately prioritize level and type of fidelity to accomplish stated goal
3) Both the user and producer should be involved in development process
4) High fidelity does not necessarily mean greater training performance benefits
5) WoZ and other techniques can maintain fidelity while increasing the breadth and depth of HMT testing scenarios, especially for capabilities which are still in development
6) Trust and workload are two important measurements in understanding how HMT can most effectively be implemented in different contexts.

5 Conclusion

We have demonstrated through a variety of testbeds the utility of presenting varying levels of appropriate levels of fidelity in human-machine teaming research. Technology available today will change rapidly and levels of fidelity will likely improve accordingly. It will therefore remain a constant challenge to balance the goals of the research with the available technologies to accurately assess human-machine teaming performance for future operations. This paper is a first step to outlining an approach to assessment of appropriate levels of fidelity in human-machine teaming research.

Acknowledgements. The authors would like to thank Cadets Jessica Broll and Makenzie Hockensmith for their contributions to this work. The views expressed in this document are the authors and may not reflect the official position of the USAF Academy, USAF, or U.S. Government. The material is based upon work supported by the Air Force Office of Scientific Research under award number 16RT0881.

References

1. Sheridan, T.B.: Adaptive automation, level of automation, allocation authority, supervisory control, and adaptive control: distinctions and modes of adaptation. IEEE Trans. Syst. Man Cybern.-Part A: Syst. Hum. **41**(4), 662–667 (2011)
2. Lyons, J.B., et al.: Comparing trust in auto-GCAS between experienced and novice air force pilots. Ergon. Des. **25**(4), 4–9 (2017)
3. Ilachinski, A.: Artificial Intelligence and Autonomy: Opportunities and Challenges (No. DIS-2017-U-016388-Final). Center for Naval Analyses, Arlington, United States (2017)
4. Kaber, D.B.: Issues in human–automation interaction modeling: presumptive aspects of frameworks of types and levels of automation. J. Cogn. Eng. Decis. Making **12**(1), 7–24 (2018)
5. Hancock, P.A.: Imposing limits on autonomous systems. Ergonomics **60**(2), 284–291 (2017)
6. Scharre, P.: Army of None: Autonomous Weapons and the Future of War. WW Norton & Company, New York (2018)

7. Kott, A., Alberts, D.S.: How do you command an army of intelligent things? Computer **50**(12), 96–100 (2017)
8. Endsley, M.R.: Autonomous Horizons: System Autonomy in the Air Force-A Path to the Future. United States Air Force Office of the Chief Scientist, AF/ST TR, 15-01 (2015)
9. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J. Cogn. Eng. Decis. Making **2**(2), 140–160 (2008)
10. Miller, C.A., Parasuraman, R.: Designing for flexible interaction between humans and automation: delegation interfaces for supervisory control. Hum. Factors **49**(1), 57–75 (2007)
11. Roscoe, S.N., Williams, A.C.: Aviation psychology (1980)
12. Munshi, F., Lababidi, H., Alyousef, S.: Low-versus high-fidelity simulations in teaching and assessing clinical skills. J. Taibah Univ. Med. Sci. **10**(1), 12–15 (2015)
13. Usoh, M., et al.: Walking> walking-in-place> flying, in virtual environments. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 359–364, July 1999
14. Alexander, A.L., Brunyé, T., Sidman, J., Weil, S.A.: From gaming to training: a review of studies on fidelity, immersion, presence, and buy-in and their effects on transfer in PC-based simulations and games. DARWARS Train. Impact Group **5**, 1–14 (2005)
15. Dion, D.P., Smith, B.A., Dismukes, P.: The Cost/Fidelity Balance: Scalable Simulation Technology-A New Approach to High-Fidelity Simulator Training at Lower Cost. MS AND T, 38-45 (1996)
16. Wong, Y.J., Steinfeldt, J.A., LaFollette, J.R., Tsao, S.C.: Men's tears: football players' evaluations of crying behavior. Psychol. Men Masc. **12**(4), 297 (2011)
17. Taylor, H.L., Lintern, G., Koonce, J.M.: Quasi-transfer as a predictor of transfer from simulator to airplane. J. Gen. Psychol. **120**(3), 257–276 (1993)
18. Taylor, H.L., Lintern, G., Koonce, J.M., Kaiser, R.H., Morrison, G.A.: Simulator scene detail and visual augmentation guidance in landing training for beginning pilots. SAE Trans. **100**, 2337–2345 (1991)
19. Flexman, R.E., Stark, E.A.: Training simulators. In: Handbook of Human Factors, vol. 1, pp. 1012–1037 (1987)
20. McClernon, C.K., McCauley, M.E., O'Connor, P.E., Warm, J.S.: Stress training improves performance during a stressful flight. Hum. Factors **53**(3), 207–218 (2011)
21. Lievens, F., Patterson, F.: The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. J. Appl. Psychol. **96**(5), 927 (2011)
22. Massoth, C., et al.: High-fidelity is not superior to low-fidelity simulation but leads to overconfidence in medical students. BMC Med. Educ. **19**(1), 29 (2019). https://doi.org/10.1186/s12909-019-1464-7
23. Salas, E., Bowers, C.A., Rhodenizer, L.: It is not how much you have but how you use it: toward a rational use of simulation to support aviation training. Int. J. Aviat. Psychol. **8**(3), 197–208 (1998)
24. Choi, W., et al.: Engagement and learning in simulation: recommendations of the Simnovate engaged learning domain group. BMJ Simul. Technol. Enhanc. Learn. **3**(Suppl 1), S23-S32 (2017)
25. Tossell, C., et al.: Human factors capstone research at the united states air force academy. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 63, no. 1, pp. 498–502. SAGE Publications, Los Angeles, November 2019
26. Bishop, J., et al.: CHAOPT: a testbed for evaluating human-autonomy team collaboration using the video game overcooked! 2. In: 2020 Systems and Information Engineering Design Symposium (SIEDS), pp. 1–6. IEEE, April 2020

27. Tanibe, T., Hashimoto, T., Karasawa, K.: We perceive a mind in a robot when we help it. PloS One **12**(7), 1–12 (2017)
28. Bartneck, C., Forlizzi, J.: A design-centred framework for social human-robot interaction. In: Proceedings of the Ro-Man 2004, Kurashiki, pp. 591–594 (2004)
29. Steinfeld, A., Jenkins, O.C., Scassellati, B.: The oz of wizard: simulating the human for interaction research. In: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, pp. 101–108, March 2009
30. Lorenz, G.T., et al.: Assessing control devices for the supervisory control of autonomous wingmen. In: 2019 Systems and Information Engineering Design Symposium (SIEDS), pp. 1–6. IEEE, April 2019
31. Tomzcak, K., et al.: Let Tesla park your Tesla: driver trust in a semi-automated car. In: 2019 Systems and Information Engineering Design Symposium (SIEDS), pp. 1–6. IEEE, April 2019
32. Tenhundfeld, N.L., de Visser, E.J., Ries, A.J., Finomore, V.S., Tossell, C.C.: Trust and distrust of automated parking in a Tesla model X. Hum. Factors **62**, 194–210 (2019). 0018720819865412
33. Tenhundfeld, N.L., de Visser, E.J., Haring, K.S., Ries, A.J., Finomore, V.S., Tossell, C.C.: Calibrating trust in automation through familiarity with the autoparking feature of a Tesla model X. J. Cogn. Eng. Decis. Making **13**(4), 279–294 (2019)
34. Haring, K., Nye, K., Darby, R., Phillips, E., de Visser, E., Tossell, C.: I'm not playing anymore! A study comparing perceptions of robot and human cheating behavior. In: Salichs, M., et al. (eds.) ICSR 2019. LNCS (LNAI), vol. 11876, pp. 410–419. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-35888-4_38
35. Peterson, J., Cohen, C., Harrison, P., Novak, J., Tossell, C., Phillips, E.: Ideal warrior and robot relations: stress and empathy's role in human-robot teaming. In: 2019 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, pp. 1–6 (2019)
36. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors **46**(1), 50–80 (2004)
37. de Visser, E.J., et al.: Towards a theory of longitudinal trust calibration in human–robot teams. Int. J. Soc. Robot. **12,** 459–478 (2020). https://doi.org/10.1007/s12369-019-00596-x
38. Robinette, P., Li, W., Allen, R., Howard, A.M., Wagner, A.R.: Overtrust of robots in emergency evacuation scenarios. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 101–108. IEEE, March 2016
39. Wagner, A.R., Borenstein, J., Howard, A.: Overtrust in the robotic age. Commun. ACM **61**(9), 22–24 (2018)
40. Okamura, K., Yamada, S.: Adaptive trust calibration for supervised autonomous vehicles. In: Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, pp. 92–97, September 2018
41. Berka, C., et al.: EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviat. Space Environ. Med. **78**(5), B231–B244 (2007)
42. Ekman, P., Friesen, W.V.: Facial Action Coding Systems. Consulting Psychologists Press, Palo Alto (1978)
43. Walliser, J.C., de Visser, E.J., Wiese, E., Shaw, T.H.: Team structure and team building improve human-machine teaming with autonomous agents. J. Cogn. Eng. Decis. Making **13**(4), 258–278 (2019)
44. Demir, M., McNeese, N.J., Cooke, N.J.: Team situation awareness within the context of human-autonomy teaming. Cogn. Syst. Res. **46**, 3–12 (2017)

45. Phillips, E., Ososky, S., Grove, J., Jentsch, F.: From tools to teammates: toward the development of appropriate mental models for intelligent robots. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 55, no. 1, pp. 1491–1495. SAGE Publications, Los Angeles, September 2011
46. Garreau, J.: Bots on the ground. Washington Post **6** (2007)
47. Wen, J., Stewart, A., Billinghurst, M., Dey, A., Tossell, C., Finomore, V.: He who hesitates is lost (… in thoughts over a robot). In: Proceedings of the Technology, Mind, and Society, pp. 1–6 (2018)
48. Wen, J., Stewart, A., Billinghurst, M., Tossell, C.: Band of brothers and bolts: caring about your robot teammate. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1853–1858. IEEE, October 2018
49. Tomova, L., Majdandžić, J., Hummer, A., Windischberger, C., Heinrichs, M., Lamm, C.: Increased neural responses to empathy for pain might explain how acute stress increases prosociality. Soc. Cogn. Affect. Neurosci. **12**(3), 401–408 (2017)
50. National Adult Spelling Bee Practice. https://www.vocabulary.com/lists/144082. Accessed 23 Feb 2020
51. Phillips, E., Zhao, X., Ullman, D., Malle, B.F.: What is human-like? Decomposing robots' human-like appearance using the Anthropomorphic roBOT (ABOT) Database. In: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, pp. 105–113, February 2018
52. Kim, B., Bruce, M., Brown, L., de Visser, E., Phillips, E.: A comprehensive approach to validating the uncanny valley using the Anthropomorphic RoBOT (ABOT) database. In: 2020 Systems and Information Engineering Design Symposium (SIEDS), pp. 1–6, April 2020
53. Haring, K.S., et al.: Conflict mediation in human-machine teaming: using a virtual agent to support mission planning and debriefing. In: 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 1–7. IEEE, October 2019
54. Bellas, A., et al.: Rapport building with social robots as a method for improving mission debriefing in human-robot teams. In: 2020 Systems and Information Engineering Design Symposium (SIEDS), pp. 160–163. IEEE, April 2020
55. Haring, K.S., et al.: Robot authority in human-machine teams: effects of human-like appearance on compliance. In: Chen, J., Fragomeni, G. (eds.) HCII 2019. LNCS, vol. 11575, pp. 63–78. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21565-1_5
56. Giubilini, A., Savulescu, J.: The artificial moral advisor. The "Ideal Observer" meets artificial intelligence. Philos. Technol. **31**(2), 169–188 (2018)
57. Malle, B.F.: Integrating robot ethics and machine morality: the study and design of moral competence in robots. Ethics Inf. Technol. **18**(4), 243–256 (2016). https://doi.org/10.1007/s10676-015-9367-8
58. Savulescu, J., Maslen, H.: Moral enhancement and artificial intelligence: moral AI?. In: Romportl, J., Zackova, E., Kelemen, J. (eds.) Beyond Artificial Intelligence. TIEI, vol. 9, pp. 79–95. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-09668-1_6
59. Coovert, M.D., Arbogast, M.S., de Visser, E.J.: The cognitive Wingman: considerations for trust, humanness, and ethics when developing and applying AI systems. In: McNeese, S., Endsley (eds.) Handbook of Distributed Team Cognition. CRC Press Taylor & Francis, Boca Raton (in press)
60. Costa, A., et al.: Your morals depend on language. PLoS One **9**(4), e94842 (2014)
61. Greene, J.D., Morelli, S.A., Lowenberg, K., Nystrom, L.E., Cohen, J.D.: Cognitive load selectively interferes with utilitarian moral judgment. Cognition **107**(3), 1144–1154 (2008)
62. Sütfeld, L.R., Gast, R., König, P., Pipa, G.: Using virtual reality to assess ethical decisions in road traffic scenarios: applicability of value-of-life-based models and influences of time pressure. Front. Behav. Neurosci. **11**, 122 (2017)

63. Tinghög, G., et al.: Intuition and moral decision-making – the effect of time pressure and cognitive load on moral judgment and altruistic behavior. PLoS One **11**(10), e0164012 (2016)
64. Cook, M.L.: The Moral Warrior: Ethics and Service in the U.S. Military. SUNY Press, Albany (2004)
65. Williams, T., Zhu, Q., Wen, R., de Visser, E.J.: The confucian matador: three defenses against the mechanical bull. In: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, pp. 25–33, March 2020
66. Jackson, R.B., Williams, T.: Language-capable robots may inadvertently weaken human moral norms. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 401–410 (2019)
67. Rosemont Jr, H., Ames, R.T.: Confucian Role Ethics: A Moral Vision for the 21st Century? Vandenhoeck & Ruprecht, Göttingen (2016)
68. Groom, V., Nass, C.: Can robots be teammates?: Benchmarks in human–robot teams. Interact. Stud. **8**(3), 483–500 (2007)
69. Murphy, R.R.: Disaster Robotics. MIT Press, Cambridge (2014)
70. Ho, N.T., Sadler, G.G., Hoffmann, L.C., Lyons, J.B., Johnson, W.W.: Trust of a military automated system in an operational context. Milit. Psychol. **29**(6), 524–541 (2017)
71. Kim, B., et al.: How early task success affects attitudes toward social robots. In: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, pp. 287–289, March 2020
72. Schellin, H., et al.: Man's new best friend? Strengthening human-robot dog bonding by enhancing the Doglikeness of Sony's Aibo. In: 2020 Systems and Information Engineering Design Symposium (SIEDS), pp. 1–6. IEEE, April 2020